

Un système collectif d'utilisation d'un grand ensemble de classifieurs sur le Cloud pour la classification de Big Data

Rabah Mazouzi*, Cyril de Runz**, Herman Akdag*

*LIASD, Université Paris 8, 2 rue de la Liberté - 93526 Saint-Denis cedex
rabah@ai.univ-paris8.fr, akdag@ai.univ-paris8.fr
<http://www.ai.univ-paris8.fr/>

**CRESTIC, IUT de Reims, Chemin des Rouliers CS30012 51687 REIMS CEDEX 2
cyril.de-runz@univ-reims.fr
<http://crestic.univ-reims.fr/>

Résumé. Au vu de l'évolution des volumes de données (Big Data) et des problématiques associées (vélocité, variété et véracité), nous proposons dans cet article la conception d'un nouveau système collectif d'utilisation massive d'ensemble de classifieurs pour les Big Data sur le Cloud. Nous combinons les avantages de la labellisation par consensus entre plusieurs décisions de classifieurs distribués sur le Cloud avec l'utilisation du paradigme Map/Reduce pour l'apprentissage des modèles par chacun des classifieurs. Pour cela, nous considérons un réseau de classifieurs déployé sur le Cloud. Par l'intermédiaire des Mappers, nous répartissons les données d'apprentissage sur les différents nœuds (classifieurs) tandis que les Reducers lancent la phase d'apprentissage et retournent le modèle du classifieur ainsi qu'un indicateur de performance à optimiser. Ensuite, pour chaque donnée qui arrive, quel que soit le nœud du réseau sur lequel elle arrive, le nœud labellise la donnée et demande à ces voisins d'en faire tout autant. Ils forment ainsi un ensemble de classifieurs. Enfin, à l'aide d'un vote majoritaire pondéré, le nœud questionné renvoie la décision finale. Ainsi, plus le voisinage est étendu, plus la performance cherchée s'améliore. Cependant, il faut limiter cette extension car sinon nous n'obtenons plus des temps de traitements compatibles avec les Big Data.

1 Introduction

Généralement, les algorithmes de classification utilisent, pour la phase d'apprentissage, des ensembles de données limités en taille et en nature. La problématique de la classification prend une autre dimension avec des données très volumineuses (Big Data), notamment à cause du volume et de la variété des données, ainsi que de la vitesse de réponse du système. Pour pallier aux problèmes liés à la classification des Big Data, le partitionnement des données sur un nombre élevé de classifieurs de nature diverse, constitue, selon nous, une solution idéale.

De nos jours, de nombreuses ressources sont disponibles et mises à disposition dans l'objectif de mettre en place des solutions autrefois très coûteuses et peu accessibles. Ainsi, le